Using Coreference Links to Improve Spanish-to-English Machine Translation

Lesly Miculicich Werlen and **Andrei Popescu-Belis**
Idiap Research Institute
Rue Marconi 19, CP 592
1920 Martigny, Switzerland
{lmiculicich,apbelis}@idiap.ch

Abstract

In this paper, we present a proof-of-concept of a coreference-aware decoder for document-level machine translation. We consider that better translations should have coreference links that are closer to those in the source text, and implement this criterion in two ways. First, we define a similarity measure between source and target coreference structures, by projecting the target ones onto the source and reusing existing coreference metrics. Based on this similarity measure, we rerank the translation hypotheses of a baseline system for each sentence. Alternatively, to address the lack of diversity of mentions in the MT hypotheses, we focus on mention pairs and integrate their coreference scores with MT ones, resulting in post-editing decisions. Experiments with Spanish-to-English MT on the AnCora-ES corpus show that our second approach yields a substantial increase in the accuracy of pronoun translation, with BLEU scores remaining constant.

1 Introduction

Considering entire texts for machine translation, rather than separate sentences, has the potential to improve the consistency of the translations. In this paper, we focus on coreference links, which connect referring expressions that denote the same entity within or across sentences. As perfect translations should provide the reader the same understanding of entities as the source texts, we propose to use the similarity of coreference links between a source text and its translation as a criterion to improve translation hypotheses. This information should be beneficial to the translation of pronouns, which often depends on the properties of their antecedent, but should also ensure lexical consistency in the translation of coreferent nouns.

We provide here the first proof-of-concept showing that the coreference criterion can lead to measurable improvements in the translation of referring expressions such as noun phrases and, especially, pronouns – in the case of Spanish to English machine translation (MT). To implement this criterion, we first need to compute coreference links in the source and target texts. We then compare two approaches: either computing a global coreference score by comparing the links and using it to rerank the hypotheses of an MT system; or integrating mention-pair scores from a coreference resolution system with MT scores, and post-editing each mention to maximize these scores. These approaches are presented in the paper as follows, preceded by an overview of related work on coreference and anaphora resolution and translation (Section 2).

In Section 3, for computing source and target-side coreference links, we take advantage of gold standard coreference links on the Spanish AnCora-ES corpus, and use the Stanford Coreference Resolution system on the English MT output. These are used for both coreference-aware MT methods that we present. In Section 4, we compare coreference links globally by projecting the referring expressions (mentions) from target to source texts, and measuring similarity by using existing coreference resolution metrics (MUC, B3, CEAF). As a sanity check, in Section 4.2, we show that better translations, in the sense of higher BLEU scores, exhibit higher coreference similarity scores as well. Global coreference similarity is then used in Section 4.3 as a constraint to rerank hypotheses of the Moses MT decoder. Alternatively, as the top MT hypotheses do not vary enough in terms of mentions, we propose in Sec-

tion 5 a different method, which focuses on the translation variants of the mentions, and post-edits them using information from coreference chains in the source text. Finally, the results presented in Section 6 show that the second method increases mainly the accuracy of pronoun translation from Spanish to English, while obtaining BLEU scores similar to those of the MT baseline.

2 Related Work

2.1 Coreference Resolution and Evaluation

Coreference resolution is the task of grouping the expressions that refer to the same entity in a text. This task includes two stages: mention identification, and coreference resolution. The first stage is usually based on part-of-speech annotation and named-entity recognition. Candidate mentions are usually noun phrases, pronouns, and named entities (Lee et al., 2011). Coreference resolvers follow three main approaches: pairwise, re-ranking, and clustering. Pairwise resolvers perform a binary classification, predicting if two mentions refer to the same entity or not. This assumes strong independence of mentions and does not utilize features of the entire entity (Bengtson and Roth, 2008). The second approach lists a set of candidate antecedents for each mention that are simultaneously considered to find the best match. Interpolation between the best and worse candidate is considered (Wiseman et al., 2015; Bengtson and Roth, 2008). Finally, the clustering approach considers the features of a complete cluster of mentions to decide whether a mention belongs or not to a cluster (Clark and Manning, 2015; Fernandes et al., 2012).

Coreference resolution is typically evaluated in comparison with a gold-standard annotation (Popescu-Belis, 1999; Recasens and Hovy, 2011). The main metrics used for evaluation are MUC (Vilain et al., 1995), which counts the minimum number of links between mentions to be inserted or deleted to map the evaluated document to the gold-standard. The B^3 measure (Bagga and Baldwin, 1998) computes precision and recall for all mentions of a document, while CEAF (Luo, 2005) computes them at the entity level. BLANC (Recasens and Hovy, 2011) makes use of the Rand Index, an algorithm for the evaluation of clustering. These metrics are implemented in the scorer for CoNLL 2012 (Pradhan et al., 2014) and the SemEval 2013 one (Màrquez et al., 2013).

2.2 Coreference-Aware Machine Translation

Despite the numerous coreference and anaphora resolution systems designed in the past decades (Mitkov, 2002; Ng, 2010), the interest in using them to improve pronoun translation has only recently emerged (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Guillou, 2012; Luong et al., 2015). The still limited accuracy of coreference resolution may explain its restricted use in MT, although, it has long been known that some pronouns require knowledge of the antecedent for correct translation. For instance, Le Nagard and Koehn (2010) trained an English-French translation model on an annotated corpus in which each occurrence of the English pronouns *it* and *they* was annotated with the gender of its antecedent on the target side. Their system correctly translated 40 pronouns out of the 59 that they examined, but did not outperform the MT baseline. Recently, a model for MT decoding proposed by Luong (2016; 2017) used in a probabilistic way several features of the antecedent candidates (especially the gender, number and humanness values), and demonstrated some improvement on pronouns.

Two shared tasks on pronoun-focused translation have been recently organized. The improvement of pronoun translation was only marginal with respect to a baseline SMT system in the 2015 shared task (Hardmeier et al., 2015), while the 2016 shared task (Guillou et al., 2016) was only aiming at pronoun prediction given source texts and lemmatized reference translations. Some of the best systems developed for these tasks avoided, if fact, the direct use of anaphora resolution. For example, Callin et al. (2015) designed a classifier based on a feed-forward neural network, which considered as features the preceding nouns and determiners along with their parts-of-speech tags. The winning systems of the 2016 task used deep neural networks: Luotolahti et al. (2016) and Dabre et al. (2016) summarized the preceding and following contexts of the pronoun to predict and passed them to a recurrent neural network. To the best of our knowledge, we present here the first proof-of-concept that coreference links across noun phrases and pronouns can serve to improve statistical MT.

3 Coreference Resolution for MT

A principle of translation is that the information conveyed in a document should be preserved in

Source	Human Translation	Machine Translation
La película narra la historia de [un joven parisiense]$_{c_1}$ que marcha a Rumanía en busca de [una cantante zíngara]$_{c_2}$, ya que [su]$_{c_1}$ fallecido padre escuchaba siempre [sus]$_{c_2}$ canciones.	The film tells the story of [a young Parisian]$_{c_1}$ who goes to Romania in search of [a gypsy singer]$_{c_2}$, as [his]$_{c_1}$ deceased father use to listen to [her]$_{c_2}$ songs.	The film tells the story of [a young Parisian]$_{c_1}$ who goes to Romania in search of [a gypsy singer]$_{c_2}$, as [his]$_{c_2}$ deceased father always listened to [his]$_{c_2}$ songs.
Pudiera considerarse un viaje fallido, porque [$\emptyset$]$_{c_1}$ no encuentra [su]$_{c_1}$ objetivo, pero el azar [le]$_{c_1}$ conduce a una pequea comunidad...	It could be considered a failed journey, because [he]$_{c_1}$ does not find [his]$_{c_1}$ objective, but the fate leads [him]$_{c_1}$ to a small community...	It could be considered [a failed trip]$_{c_3}$, because [it]$_{c_3}$ does not find [its]$_{c_3}$ objective, but the chance leads $\emptyset$ to a small community...

Table 1: Comparison of coreference chains in the Spanish source vs. English human and machine translations. English chains were obtained with the Stanford coreference resolver (Manning et al., 2014). The chains are numbed $c_1, c_2, \ldots$ and are also color-coded. The void symbol $\emptyset$ indicates a correct null subject pronoun in Spanish, and an incorrect object pronoun dropped by the MT system. The third coreference chain (c_3) in the MT output is erroneous.

its translation. Here, we focus on the referential information, i.e. the coreference links between mentions. If we apply coreference resolution to a source text and to a faithful translation of it, then the grouping of mentions should be identical. We thus formulate the following criterion for MT: *better translations should have coreference links that are more similar to the source.*

Table 1 illustrates the above criterion on an example of Spanish-to-English translation, extracted from the AnCora-ES corpus (Recasens and Martí, 2010) [1], with source coreference chains coming from the AnCora-ES annotations. The automatic translation comes from a commercial online MT system, while the human translation was done by the authors of this paper. The Stanford Coreference Resolution system (Manning et al., 2014) [2] was applied to both translations, and the resulting coreference chains are indicated in the table with numbers and colors. We observe that the chains in the human translation match well those in the source, but this is less the case for the automatic translation, in particular due to wrong pronoun translations. Although the MT output is still understandable, this requires more time than with the human translation, due to the wrong set of coreference links inferred by the reader.

In what follows, we will implement a proof-of-concept coreference-aware MT system for Spanish-to-English translation. This pair is particularly challenging because Spanish is a pro-drop language, so that an MT system must not only select the correct translation of pronouns, but it must also generate English pronouns from Spanish null ones. In this study, in order to avoid introducing errors made by the coreference resolution system, we will always use on the source side the gold-standard coreference annotation from AnCora-ES (Recasens and Martí, 2010), which was used in the SemEval-2010 task 1: "Coreference resolution in multiple languages" (Verhagen et al., 2010) [3]. As our proposal does not require specific training on coreference-annotated data, AnCora-ES will be used for testing only.

On the target side, as coreference resolution must be performed for each translation hypothesis, we must use an automatic system. One advantage of the Spanish-to-English direction is that English coreference resolution systems have been studied and developed for a long time, more than any other language, thus keeping coreference errors to a minimum. We use here the statistical coreference resolution system proposed by Clark and Manning (2015) (Stanford Statistical Coreference Resolution). Moreover, to obtain pairwise mention scores, needed in Section 5, we use the pairwise classifier implemented in the Stanford CoreNLP toolkit (Manning et al., 2014) [4].

[1] http://clic.ub.edu/corpus/

[2] http://nlp.stanford.edu/projects/coref.shtml

[3] http://stel.ub.edu/semeval2010-coref/

[4] http://stanfordnlp.github.io/CoreNLP/index.html

4 Using Coreference Similarity to Rerank MT Hypotheses

4.1 Measuring Coreference Similarity

After applying coreference resolution to the source and a candidate translation, we need to compare the sets of coreference links, with the source playing the role of the ground-truth or gold-standard. Traditional metrics for evaluating coreference resolution could be used, but they have been designed to compare texts in the same language, and not across different languages, which raises difficulties for matching the referring expressions (i.e. mentions, or markables).

We propose to project the mentions of the target text back to the source text, so that each word in the source is aligned with its corresponding translation (one or more words). This alignment can be obtained directly from the Moses MT system (see start of Section 4.3).

There is not always a one-to-one word correspondence between the words in the source and target sentences, and word order also differs. Thus, we apply the following heuristic to improve the cross-language mapping of the mentions. As through word-alignment the words that comprise the mentions may have changed order in the translation, we take the first and last words in the target side, aligned to any word of the mention in the source, and we assume that all words in between are also part of the mention. The null pronouns are transfer to the next immediate verb, and we refine the alignment to be sure these verbs are aligned to the generated pronoun in the target.

Once the target mentions are mapped to the source, we apply the MUC, B^3 and CEAF-m coreference similarity metrics from the CoNLL 2012 scorer (see Section 2.1) between the source document d_s and the projected target one d_t. To mitigate individual variations, we use the average of the three scores at the similarity criterion and note it $C_{sim}(d_t, d_s)$. We did not include BLANC in this pool based on initial experiments that showed that its rate of variation was much higher than the other three metrics.

4.2 Validating the Relationship between Coreference and Translation Quality

To validate the insight that better translations correlate with better coreference similarity scores, we present in Table 2 the MUC, B^3 and CEAF scores of a human translation vs. two systems: the Moses baseline phrase-based MT system used below and an online commercial MT system using neural networks. The source is a set of documents with ca. 3.5 thousand words with gold-standard coreference annotation from AnCora-ES. The English translation was done by the authors of the paper. On the target side, we applied the Stanford automatic coreference resolution system (Manning et al., 2014).

By definition, the best translation is made by the human. Then, according BLEU score measured on the same set of documents, the second best translation is made by the commercial MT with 49.4, and the last one by the baseline MT with 43.7. We observe that the coreference scores also decrease in this order, and they decrease consistently for the three evaluation metrics. These results thus support the principle that translation quality and coreference similarity are correlated. We will now show how to use this principle to improve translation quality.

Metric	Translation	Recall	Prec.	F1
MUC	Human	31	46	37
	Commercial MT	21	38	28
	Baseline MT	18	33	23
B^3	Human	24	49	32
	Commercial MT	20	38	26
	Baseline MT	17	40	24
CEAF	Human	41	40	41
	Commercial MT	34	39	36
	Baseline MT	32	35	33

Table 2: Coreference similarity scores (%) between source and target texts for different translations. The scores increase with the quality of translations.

4.3 Reranking MT Hypotheses

We propose to use the document-level coreference similarity score C_{sim} defined above to rerank for each sentence the n-best hypotheses of an MT system. The coreference similarity is not measured individually for each sentence, but at the document level. Our goal is to find a combination of translations that optimizes this global score.

For this purpose, we use the Moses toolkit to build a phrase-based statistical MT system (Koehn et al., 2007), with training data from the translation task of the WMT 2013 workshop (Bojar et al., 2013). The English-Spanish training set consists of 14 million sentences, with approximately 340 million tokens. The tuning set is the *News Test 2010-2011* one, with ca. 5,500 sentences and

almost 120k tokens. We built a 4-gram language model from the same training data augmented by ca. 5,500 sentences monolingual data from *News Test 2015*. Our baseline system has a BLEU score of 30.8 on the *News Test 2013* with 3,000 sentences.

We thus model the problem as follows. A translated document d_t is represented as an array of translations $d_t = (s^1, s^2, ..., s^M)$, where each sentence can be selected from a list of n-best translation hypotheses $s^i \in \{s^i_1, s^i_2, ..., s^i_N\}$. The objective is to select the best combination of hypotheses based on their coreference similarity C_{sim} with the source, i.e.:

$$\arg\max_{h_1,h_2,..,h_M} C_{sim}((s^1_{h_1}, s^2_{h_2}, ..., s^m_{h_M}), d_s)$$

To limit the decrease of sentence-level translation scores when optimizing the document-level objective, we keep track of the former and select the sentences with the best translation scores if they lead to the same C_{sim}.

This combinatorial problem is expensive, so we try to reduce the search space to allow reasonable performance. First, we filter out candidate sentences. In this approach, the important variations in translation are the mentions, thus sentences are modeled as sets of mentions and duplicate sets are filtered out. Second, we apply beam search optimization. Based on the fact that the first mentions of entities usually contain more information than the next ones, the beam search starts from the first sentence and aggregates at each step the translation hypothesis with the highest similarity scores with the preceding ones.

We foresee several limitations of this approach. First, with a sentence containing several mentions, there is no guarantee that the n-best hypotheses include a combination of mention translations that optimize all mentions as the same time. What is worse, the correct translation of a given mention may not be present at all among the n-best hypotheses, because the differences among the top hypotheses are often very small, especially when sentences are long. In order to solve these problems, we present a second approach.

5 Post-editing Mentions Based on Mention Pair and MT Scores

This approach differs from the previous one in two aspects. First, it uses hypotheses of translation of individual coreferent mentions rather than of complete sentences. This allows to optimize the translation of each mention independently, and to increase the variety of hypotheses of each mention. Second, coreference resolution is applied only in the source side. So, instead of searching for similar clustering in the target side, we try to induce it. The selection of the best translation hypothesis of a mention is based on a cluster-level coreference score. We choose the hypothesis that correlates better with other mentions in the same cluster. This method improves the performance because it uses coreference resolution only once instead of multiple times, and as shown in the experimental section, it is more effective at improving the translation of mentions.

5.1 Selecting Candidate Translations

In order to obtain the n-best translation hypotheses of the mentions, it is important to include the surrounding context in the translation, otherwise, an independent translation could lead to the construction of invalid or erroneous sentences.

We would like to have a MT system that brings hypotheses corresponding only to mentions and fix the translations of other word, in a way that we can interchange the hypotheses of one mention in the same text. Building such MT system would require a significant modification of the baseline.

As an alternative solution, we will simply perform two passes of MT. The first pass is a simple translation of the text. Then, the mentions are identified in the target text and they are replaced by their source-language version. This results into a mixed language text that will be passed a second time to the MT system, so that the system will identify and translate only the words in the source language. Nevertheless, the language and reordering models are still going to evaluate on the complete sentence. To avoid any translation of the context words (i.e. not mentions) in the second pass, we filter out from the translation table all words not corresponding to mentions.

It is important to note that we consider only the heads of mentions obtained from the parse tree (this annotation is included in AnCora corpus), in order to avoid long mentions such as the ones with subordinate clauses, and focus on the most important part of each mention.

5.2 Cluster-level Coreference Score

In this approach, we rely on the coreference resolver applied to the source side to define the clusters of mentions. Each cluster is defined as a set of mentions $c_x = \{m^i, m^j, .., m^k\}$, where each mention can be selected from a set of translation hypotheses $m^i \in \{m^i_1, m^i_2, ..., m^i_N\}$.

By definition, the mentions in a cluster represent the same entity. Thus, they have to correlate in features such as gender, number, animation, etc. In order to achieve this objective in the target side, we define a cluster-level coreference score C_{ss}. It represents the likelihood that all mentions in that cluster belong to the same entity. So, for each given cluster, we select the combination of translation hypotheses of mentions with higher cluster-level coreference score.

This combinatorial problem is expensive, therefore, it is simplified with a beam search approach. Mentions are processed one at a time. The translation hypotheses of a new upcoming mention are compared with each of the previously selected ones. Then, the combinations with lower C_{ss} are pruned. The algorithm continue in the same manner until it processes the last mention.

In order to compare two mentions, we use the mention pair scorer from (Clark and Manning, 2015). It uses a logistic classifier to assign a probability to a pair of hypotheses, which represents the likelihood that they are coreferent. The pair score is defined as follows:

$$p_{pair}(m^i_{h_i}, m^j_{h_j}) = (1 + e^{\theta^T f(m^i_{h_i}, m^j_{h_j})})^{-1}$$

where $f(m^i_{h_i}, m^j_{h_j})$ is a vector of feature functions of the mentions and θ is the vector of feature weights. Finally, we define the cluster-level coreference score C_{ss} as the product of the individual pairwise probabilities:

$$C_{ss}(c_x) = \prod_{m^i \in c_x} \prod_{m^{i \neq j} \in c_x} p_{pair}(m^i_{h_i}, m^j_{h_j})$$

We illustrate this idea with an example. Here, we have a sentence in Spanish and its translation to English. We show one coreference cluster c_1 formed by three mentions:

Source (es): *La alcaldesa de Málaga y cabeza del* $[partido]_{c_1}$ $[que]_{c_1}$ *ganó en esta ciudad, pidió a los militantes de* $[este\ partido\ político]_{c_1}$...

Target (en): *The mayor of Malaga and head of the* $[m_1]_{c_1}$ $[m_2]_{c_1}$ *won in this city, asked the militants of this* $[m_3]_{c_1}$ *to...*

In this example, the three marked mentions have the following translation hypotheses: $m_1 \in \{match, party\}$, $m_2 \in \{who, which\}$, and $m_3 \in \{political\ party\}$. We calculate the pairwise score p_{pair} of each combination and show the results in the following table.

m_1, m_2	$(match, who) = 0.03, (match, which) =$ **0.35**, $(party, who) = 0.01, (party, which) =$ **0.26**
m_1, m_3	$(match, political\ party) = 0.08$, $(party, political\ party) =$ **0.53**
m_2, m_3	$(political\ party, who) = 0.12$, $(political\ party, which) =$ **0.27**

Finally, we find that the set of translation hypotheses with the highest cluster-level coreference C_{ss} score is {'party', 'which', 'political party'}, with a score of 0.04. Intuitively, we can verify that this final combination is the best solution for the example.

5.3 Incorporating Entity and Translation Information

The proposed score guides the system to select translation hypotheses which are more likely to refer to the same entity in a cluster. In order to enhance the decision process, we include two sources of additional information: the translation frequency, that can help to decide between synonyms words by selecting the most frequently translated one; and information of the entity in the source side, which enriches the knowledge of the entity.

The information about frequency of translation can indicate how well a particular hypothesis translates the mention. Therefore, we define a translation score, T_s, at mention-level. The translation score of a hypothesis is calculated based on its relative frequency of emission by the MT system, as follows:

$$T_s(m^i_{hi}) = count(m^i_{hi}) / \sum_j count(m^i_j)$$

The information about the entity in source side can indicate how well a particular hypothesis represents it. Thus, we define a simple representation of an entity by setting relevant features such as gender, number, and animation. The features are extracted and summarized from all mentions in the cluster. This is a naive representation, and more advanced work on entity-level representations has been performed in relation to coreference resolution (Clark and Manning, 2016; Wiseman et al., 2016), which could be applied here in the future.

Having an entity representation, we define a simple scoring function which measures how well a candidate represents an entity with respect to other alternatives:

$$E_s(m^i_{hi} = f(m^i_{h_i}, \theta_{e_x}) / \sum_j f(m^i_j, \theta_{e_x})$$

where f is a linear function and θ_{e_x} are the entity features.

5.4 Combining Scores

Finally, the decision is made through the combination of the three previous scores: cluster-level coreference, translation, and entity matching. As one additional step, we adjust the coreference score to the same scale as others. $C_s = C_{ss}(m^i_{h_i}, m^j_{h_j}, ...) / \sum_{x,y,..} C_{ss}(m^i_x, m^j_y, ...)$

The final score is defined as follows:

$$\begin{aligned} C_{score}(m^i_{h_i}, m^j_{h_j}, ...) = & C_s(m^i_{h_i}, m^j_{h_j}, ...)^{\lambda_1} \times \\ & [T_s(m^i_{h_i}).T_s(m^j_{h_j})...]^{\lambda_2} \times \\ & [E_s(m^i_{h_i}).E_s(m^j_{h_j})...]^{\lambda_3} \end{aligned}$$

where $\sum_i \lambda_i = 1$ are predefined hyper-parameters of the function. The final set is given by:

$$(m^i, m^j, ...) = \underset{h_i, h_j, ..}{\arg\max} \quad C_{score}(m^i_{h_i}, m^j_{h_j}, ...)$$

These three hyper-parameters were optimized on a different subset of AnCora-ES than the one used for evaluation. The optimized values are λ_1=0.5, λ_2=0.1, and λ_3=0.4.

6 Experimental Results

The objective of our initial experiments is to measure how much coreference can improve the correct choices of translation of mentions, and impact of these choices on global translation quality. We translated 10 sample documents from the test set to serve as reference translations for evaluation.

The evaluation of global MT quality is made with the well-known BLEU *n*-gram precision metric (Papineni et al., 2002), while the evaluation of mentions, being less standardized, is performed in several ways. We reuse previous insights on pronoun translation and therefore score them with a metric that automatically computes the accuracy of pronoun translation (APT) in terms of number of pronouns that are identical vs. different from a human reference translation (Werlen and Popescu-Belis, 2016) [5].

	System		
Metric	**Baseline**	**Re-rank**	**Post-edit**
BLEU	46.5±4.3	41.7±3.9***	46.4±3.9
APT	0.35±0.07	0.40±0.10*	0.59±0.13***
ANT	0.78±0.08	0.74±0.01**	0.78±0.07

Table 3: Comparison of baseline MT and our proposals for reranking or post-editing, for three metrics. In addition to the average scores and standard deviation over the ten test documents, we indicate the statistical significance level of the difference between each of our systems and the baseline (* for 95.0%, ** for 99.0% and *** for 99.9%).

	System		
Evaluation	**Baseline**	**Re-rank**	**Post-edit**
No. '0' (wrong)	53	55	21
No. '1' (acceptable)	21	19	28
No. '2' (eq. to ref.)	115	115	140
Sum of the scores	251	249	308

Table 4: Manual evaluation of fourth randomly selected documents. The evaluation was done over nouns and pronouns.

More originally, in order to provide a complete view of the performance, we compute the "accuracy of noun translation" (ANT), by reusing the same idea as in APT to count the number of exactly matched nouns between MT and the reference translation. Finally, we perform manual evaluation by examining source mentions, as annotated over AnCora-ES, and evaluating their individual translations by the baseline MT along with the two approaches presented above (in Sections 4 vs. 5). When presented to the evaluator, the three translations of each source sentence are provided in a random order, so that the evaluator does not know to which system they belong. The evaluator assigned a score of '2' to a translation identical to the reference, '1' for translation that is different but still good or acceptable, and '0' to a wrong or unacceptable translation. To minimize the time spent on manual evaluation at this stage, one evaluator rated four test documents.

Table 3 shows the results of the experiments obtained with automatic metrics. We first calculate BLEU, APT, and ANT values at document-level, and show the values of the average and standard deviation for the three evaluated systems: base-

[5] https://gitlab.idiap.ch/lmiculicich/APT

line, and our two proposed approaches. Additionally, we show the significance levels (t-test) of the results in comparison to the baseline. The post-editing approach improves the pronoun translation quite significantly, without decreasing the overall quality of translation. This improvement is demonstrated by the rise of APT score, whereas BLUE score remains without significant change. However, the quality of the translation of nouns does not change significantly, as shown by the ANT. The re-ranking approach shows a significant increase in the quality of pronoun translation. Nevertheless, the overall quality of translation decreases significantly, as well as the quality of noun translation. These results can be explained by the limitations of this approach. The optimization was done by taking into account the correlation of mentions, but the changes were made at sentence level, and the overall quality of translation at sentence level was not considered. To address this problem, a combination of coreference similarity and translation probability for each sentence could be used in future.

Table 4 shows the results of the manual evaluation, scored as explained above, which includes nouns and pronouns together. In general, it supports the results of the automatic evaluation. Here, the post-editing approach has 32 less mentions scored as "wrong" than the baseline, 7 of them were score as "acceptable", and the rest 25 as identical to the reference. The re-ranking approach, despite the theoretical appeal of its definition, fails to improve noun and pronoun translation.

Figure 1 shows the distribution of pronouns translated by the three evaluated systems (i.e. baseline, re-ranking, and post-editing) in comparison with the reference. The number of pronouns equal to the reference increases for both proposed approaches, specially for the post-editing. The pronouns that improve the most were the third-person personal and possessive ones. Also, the translation of some of the null pronouns in the source was improved. The association with other mentions of the same entity, and the representation of the entity coming from the source side was important for this improvement.

Table 5 shows examples of translations obtained with our approaches. The translations of nouns are already good for the baseline, and the differences are in many cases due to the use of synonyms and acronyms. Still, there are source nouns that suffer from sense ambiguity, which may be improved by our method. However, this particular test set is too small and does not contain enough instances of this type to evaluate their translations with certainty.

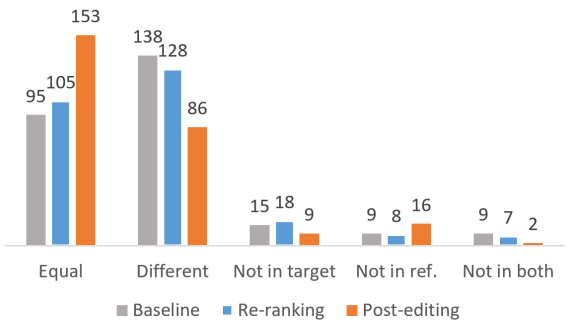

Figure 1: Pronoun translation in comparison with the reference translation: numbers of equal vs. different pronouns for the three systems, including also missing pronouns in target, reference, and both sides.

7 Conclusion

We have presented two methods for improving noun and pronoun translation based on coreference similarity of source and translated texts. While the re-ranking approach did not achieve its goals, the post-editing approach brought a significant improvement of Spanish-to-English pronoun translation. This should be confirmed, in the future, by more detailed measurements on larger data set. Also, one simplifying assumption, namely the use of ground-truth coreference annotation on the target side (here, from AnCora-ES) should be relaxed, in order to address the challenge of using automated coreference resolution on both source and target side – and thus produce a fully-automated, unrestricted MT system.

This study contributes to a growing body of research on modeling longer-range dependencies than those modeled in phrase-based or neural MT, across different sentences of a document. The Docent decoder (Hardmeier et al., 2012), which uses document-level features to improve coherence across translated sentences, could also be used in combination with the coreference similarity score, or, alternatively, neural MT could be adapted to take advantage of neural network representations of coreference information.

Correctly modified examples
S: *[Barton]$_3$, por [su]$_3$ parte , también dudó de la capacidad de [Megawati]$_2$ en [su]$_3$ [nueva tarea]$_4$.* **R:** *[Barton]$_3$, for [his]$_3$ part , also doubted [Megawati]$_2$'s ability in [her]$_2$ [new task]$_4$.* **B:** *[Barton]$_3$, for [its]$_3$ part , also doubted the capacity of Megawati in [his]$_2$ [new task]$_4$.* **P:** *[Barton]$_3$, for [his]$_3$ part , also doubted the capacity of [Megawati]$_2$ in [her]$_2$ [new task]$_4$.*
S: *... que " [parece estar]$_2$ abrumada ... críticos consideran que [no será]$_2$ capaz de hacerse con el papel de líder .* **R:** *...that " [she seems]$_2$ overwhelmed ... critics consider [she will not be]$_2$ able to take the lead role .* **B:** *... that " [appears to be]$_2$ overwhelmed ... critics believe that [it will not be]$_2$ able to take a leading role .* **P:** *...that " [she seems]$_2$ to be overwhelmed ... critics believe that [she will not be]$_2$ able to take a leading role .*
Incorrectly modified example
S: *- [Es]$_1$ iconoclasta por valenciano ? - .* **R:**: *- [Are you]$_1$ iconoclastic by Valencian ? - .* **B:** *- [Is]$_1$ an iconoclast by Valencian ? - .* **P:** *- [he is]$_1$ an iconoclast by Valencian ? - .*

Table 5: Examples of source, reference, baseline and post-edited sentences.

Acknowledgments

We are grateful for support to the Swiss National Science Foundation (SNSF) under the Sinergia MODERN project (grant n. 147653, see `www.idiap.ch/project/modern/`) and to the European Union under the Horizon 2020 SUMMA project (grant n. 688139, see `www.summa-project.eu`). We thank the CORBON anonymous reviewers for their helpful suggestions.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, volume 1, pages 563–566, Granada, Spain.

Eric Bengtson and Dan Roth. 2008. Understanding the value of features for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Honolulu, Hawaii, October. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 workshop on statistical machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, August. Association for Computational Linguistics.

Jimmy Callin, Christian Hardmeier, and Jörg Tiedemann. 2015. Part-of-speech driven cross-lingual pronoun prediction with feed-forward neural networks. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 59–64, Lisbon, Portugal, September. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2015. Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1415, Beijing, China, July. Association for Computational Linguistics.

Kevin Clark and Christopher D Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. *arXiv preprint arXiv:1606.01323*.

Raj Dabre, Yevgeniy Puzikov, Fabien Cromieres, and Sadao Kurohashi. 2016. The kyoto university cross-lingual pronoun translation system. In *Proceedings of the First Conference on Machine Translation*, pages 571–575, Berlin, Germany, August. Association for Computational Linguistics.

Eraldo Fernandes, Cícero dos Santos, and Ruy Milidiú. 2012. Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 41–48, Jeju Island, Korea, July. Association for Computational Linguistics.

Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2016. Findings of the 2016 WMT Shared Task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation*, pages 525–542, Berlin, Germany.

Liane Guillou. 2012. Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10, Avignon, France, April. Association for Computational Linguistics.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *IWSLT (International Workshop on Spoken Language Translation); Paris, France; December 2nd and 3rd, 2010.*, pages 283–289.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 2012. Document-wide decoding for phrase-

based statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1179–1190, Jeju Island, Korea, July. Association for Computational Linguistics.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal, September. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Ronan Le Nagard and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Uppsala, Sweden, July. Association for Computational Linguistics.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanfords multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34, Portland, Oregon, USA, June. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, B.C., Canada.

Ngoc Quang Luong and Andrei Popescu-Belis. 2016. Improving pronoun translation by modeling coreference uncertainty. In *Proceedings of the First Conference on Machine Translation*, pages 12–20, Berlin, Germany, August. Association for Computational Linguistics.

Ngoc Quang Luong and Andrei Popescu-Belis. 2017. Machine translation of Spanish personal and possessive pronouns using anaphora probabilities. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Valencia, Spain, April.

Ngoc Quang Luong, Lesly Miculicich Werlen, and Andrei Popescu-Belis. 2015. Pronoun translation and prediction with or without coreference links. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 94–100, Lisbon, Portugal, September. Association for Computational Linguistics.

Juhani Luotolahti, Jenna Kanerva, and Filip Ginter. 2016. Cross-lingual pronoun prediction with deep recurrent neural networks. In *Proceedings of the First Conference on Machine Translation*, pages 596–601, Berlin, Germany, August. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, MD, USA.

Lluís Màrquez, Marta Recasens, and Emili Sapena. 2013. Coreference resolution: an empirical study based on SemEval-2010 Shared Task 1. *Language Resources and Evaluation*, 47(3):661–694.

Ruslan Mitkov. 2002. *Anaphora Resolution*. Longman, London, UK.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, Uppsala, Sweden, July. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Andrei Popescu-Belis. 1999. Evaluation numérique de la résolution de la référence: Critiques et propositions. *TAL: Traitement automatique des langues*, 40(2):117–146.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland, June. Association for Computational Linguistics.

Marta Recasens and Eduard Hovy. 2011. BLANC: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Marta Recasens and M Antònia Martí. 2010. Ancora-co: Coreferentially annotated corpora for spanish and catalan. *Language Resources and Evaluation*, 44(4):315–345.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 Task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden, July. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Conference on Message Understanding*, pages 45–52, Columbia, MD, USA.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2016. Validation of an automatic metric for the accuracy of pronoun translation (APT). Technical report, Idiap.

Sam Wiseman, Alexander M. Rush, Stuart Shieber, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1416–1426, Beijing, China, July. Association for Computational Linguistics.

Sam Wiseman, Alexander M Rush, and Stuart M Shieber. 2016. Learning global features for coreference resolution. *arXiv preprint arXiv:1604.03035*.